图书情报工作
第 64 卷 第 18 期 2020 年 9 月

词向量语义扩展技术在图书馆智能咨询系统的应用与实现

■ 张乐
东南大学图书馆 南京 211189

摘 要:[目的/意义]针对目前自动问答系统在语义扩展方面存在的缺陷,提出一种基于词向量的语义扩展技术,设计并实现一个图书馆的智能咨询系统。[方法/过程]使用基于 Word2vec 词向量语义扩展技术结合中文分词、共现词匹配技术设计智能问答引擎,结合协同办公的管理理念,实现图书馆智能咨询系统的构建,并对系统的运行数据进行统计分析。[结果/结论]该系统在工作时间、咨询效果和后台管理上较好地满足设计需求,为图书馆智能化信息咨询系统建设提供参考。

关键词:词向量 语义扩展 Word2vec 智能咨询 图书馆
分类号:G250.7
DOI:10.13266/j.issn.0252-3116.2020.18.014

1 引言

为了更好地为用户提供信息咨询服务,图书馆为用户提供了基于网络的多种服务方式,如虚拟咨询系统、自动问答系统等,因为语义扩展方面的缺陷,导致信息咨询的效果不佳。随着人工智能自然语言处理技术中的词向量技术的发展,为图书馆智能咨询系统的设计和构建提供了新的思路。本文通过对基于 Word2vec 词向量语义扩展的智能咨询系统的构建,以及系统运行使用的分析,为国内图书馆智能咨询系统的建设提供参考。

2 研究综述

随着数字化技术和网络技术的应用,作为图书馆核心价值的信息咨询服务也有了跨越性的发展[1],传统的面对面的咨询模式,逐渐被网络数字参考咨询所取代[2]。在咨询服务应答模式方面,早期的数字咨询使用实时交流工具以人工方式开展咨询服务,如视频会议软件、Twitter 或 Facebook 等[3]。随着人工智能技术的快速发展,智能化信息咨询系统的应用在图书馆界逐渐兴起。在国外,汉堡大学图书馆最早使用智能问答系统,试图解决传统的人工服务应答效率低、响应缓慢的问题。在国内,清华大学图书馆[4]、上海交通大学图书馆[5]、南京大学图书馆[6]、北京工商大学图书馆[7]、西安交通大学图书馆[8]等,先后使用不同的平台和技术构建了各自的自动问答系统。但大多数的系统还处于改进和测试阶段,很少能够进入到图书馆的真实场景中应用。

图书馆智能信息咨询系统的基本流程是在接受到用户提出的问题时,首先分析用户所提出的问题,抽取其中关键词,然后在已有的语料库或者知识库中进行检索、匹配,将获取的答案反馈给用户的过程。早期的自动问答系统中应用了基于关键词的检索模式,包括问题分析、关键词提取、信息检索、答案验证等过程[9]。采用的是关键词直接与答案的匹配的模式,而在实际的应用中,由于中文的语义往往可以用多个中文文本或字串来表示,而数据库中的数据和关键词又是以独立的形式存在,没有相互的关联。所以这种基于单一关键词匹配的模式由于缺乏对自然语言的同义词语义的扩展能力,导致自动问答的答案匹配率很低。此后的研究发现在检索过程中,利用基于同义词典比对的语义扩展模式,能够有效提高中文检索的答案匹配率。即通过将问题关键词和同义词典中的词比对,提取出相关词的语义扩展方式。随后又有学者尝试应用诸如知识本体与关联数据[10]和知识图谱等知识组织模式,对数据库的数据进行关联优化,为信息检索提供基于语义的理解机制[11]。这些方式的问题是建设和维护非常复杂,在数据库较大的情况下运行的效率较低。

作者简介:张乐(ORCID:0000-0001-6099-5802),馆员,硕士,E-mail:tim0894@hotmail.com。
收稿日期:2020-03-06 修回日期:2020-04-03 本文起止页码:126-136 本文责任编辑:王传清

张乐. 词向量语义扩展技术在图书馆智能咨询系统的应用与实现[J]. 图书情报工作,2020,64(18):126-136.

随着自然语言处理中的词向量技术的发展,为语义扩展提供了较好的解决方法。词向量技术是为了使用数学模型来表示自然语言的词和其相对应的向量,并且量化和分类语言项之间的语义相似性而发明的,先后出现了 n-gram、神经网络、Word2vec 等基于统计的自然语言词向量模型,它们的特点是通过对语料的训练,能够不断优化匹配效果,目前广泛应用于语义相似度计算、机器翻译、文本匹配[12]等自然语言处理方面。

3 基于 Word2vec 词向量语义扩展

智能问答系统的主要难点在于实现准确识别用户咨询的问题并返回合适的答案,对从问题中抽取出的关键词进行语义扩展,是提高答案匹配效果的关键。基于 Word2vec 的词向量语义扩展技术可以很好地解决这个问题。Word2vec 的核心思想是通过词的上下文得到词的向量化表示,利用训练样本进行训练与学习,将语句中的词语映射成多维的词向量,通过向量之间的距离来判断词语之间的相似程度。它的优势是不需要对样本数据进行复杂处理,就可以直接进行词向量训练。基于这样的特点,可以方便将图书馆信息咨询服务中累计的有效问答都添加为训练样本,通过持续的向量训练提升词向量的准确性,而不需要人工干预。基于 Word2vec 词向量语义扩展的过程首先是对样本库进行词向量的训练,获得词向量表。然后利用学习训练过后的词向量表,找到查询词与扩展词之间的余弦值,从而判断他们之间的相似度。在获得查询结果后,设置一定的阈值,若大于设置的阈值判断为相似词,将此词作为查询词的扩展词,放入扩展后的词集中,为后续的问答匹配做准备。

本文在智能咨询系统中应用 Word2vec 词向量技术结合中文分词、共现词匹配等技术设计实现智能问答引擎,以解决目前自动问答系统的语义扩展方面的缺陷,提高图书馆信息咨询系统的使用效果。据调查,目前国内尚未发表过应用该技术构建图书馆智能咨询系统的相关研究。

4 系统设计和相关技术

东南大学图书馆根据智能咨询系统的需求、整体业务流程、总体技术架构,进行了基于 Word2vec 的智能咨询引擎与维护管理平台的设计。笔者作为该项目负责人,全程参与了系统的需求、架构、设计、测试和运行维护工作。本文是对此项目的相关技术应用和系统使用的梳理和经验总结。

4.1 图书馆智能咨询系统的需求分析

东南大学图书馆在信息服务的网络化和智能化工作中,先后使用了虚拟咨询、qq 以及自动问答机器人等系统和工具,在应用过程中发现一些需要解决的问题:①基于关键词检索的问答机器人系统对问题语义扩展能力的欠缺,导致咨询答案不准确。②基于人工的信息咨询服务在服务时间上不能满足读者的需求。③重复的事务性问题占咨询问题的比重很大,如馆舍位置、服务条款、工作时间等问题,导致咨询馆员工作量过大。④图书馆用户专业性的咨询内容需要多部门合作来解答,需要便捷的内部协同处理工作模式[13]来提高应答的准确性和时效性。基于以上问题,东南大学图书馆智能咨询系统功能需求主要归纳为以下两方面:

4.1.1 用户咨询需求

智能咨询系统是以微信端、网页端等多终端展现方式,24 小时为在校师生提供图书馆信息咨询、馆藏书目检索等服务。系统支持用户以自然语言、关键词等方式进行咨询提问。基于 Word2vec 词向量语义扩展,结合中文分词、共现词答案匹配等技术实现智能问答引擎,分析用户的提问,给出关联推荐问题答案。在智能问答引擎无法回答师生问题时,提供人工服务对接。对于未解答的用户问题,管理人员在后台回答并提交至智能咨询系统后,支持对问题的推送功能。

4.1.2 后台管理需求

对于后台系统的操作人员,提供基于协同工作的分级分角色的管理功能,不同的权限提供不同后台操作功能。如学科馆员、学生馆员、问答知识库管理员以及运维人员等。支持根据不同维度进行相关数据的统计分析,并根据用户需求制作统计报表,进行可视化展示。拥有操作权限的管理员可对知识库进行添加,修改新的问题答案,支持单条或者批量进行数据更新操作,并同步到相关数据库。

4.2 图书馆智能咨询系统的业务流程

图书馆智能咨询系统(以下简称"本系统")的业务流程是:用户通过系统发送咨询问题以后,系统通过智能问答引擎对问题进行相关处理和问答的检索,如果已有答案则返回用户,如果没有匹配答案则可转为人工服务交由后台的维护和管理平台人工处理。咨询馆员通过维护和管理平台完成人工服务、协同工作和系统管的工作。如图 1 所示:

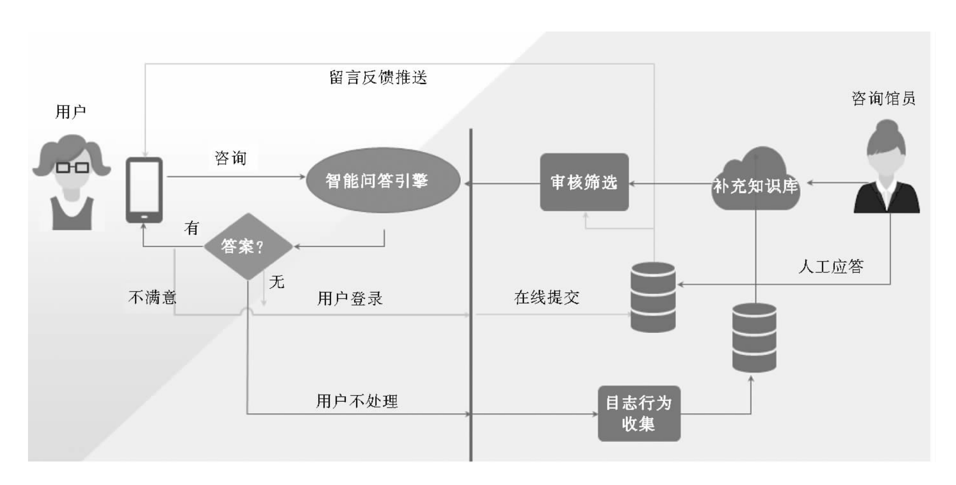


图 1　智能咨询系统业务整体流程

4.3　图书馆智能咨询系统的技术架构

图书馆智能咨询系统的架构设计(见图 2)采用了基于分层设计的思想,将智能信息咨询系统所需要提供的服务,按照功能划分为数据层、业务逻辑层、应用层和表示层。每一层子系统与上下层的子系统通过特定的子系统访问接口进行交互,不同层次的子系统均提供相应的接口,如数据层除了对智能问答的业务数据库的存储外,还提供了多种数据接口,便于支持图书馆其他业务模块的数据源的接入。在子系统的内部采用了模块化设计,各个模块相对独立,可根据未来图书馆信息咨询服务的需求灵活添加,如电子资源的检索、查证查引等应用层模块。针对图书馆用户信息获取习惯,系统提供了基于网页和微信的多终端的服务接口,同时为本馆将要引进的实体机器人的接入预留了接口。

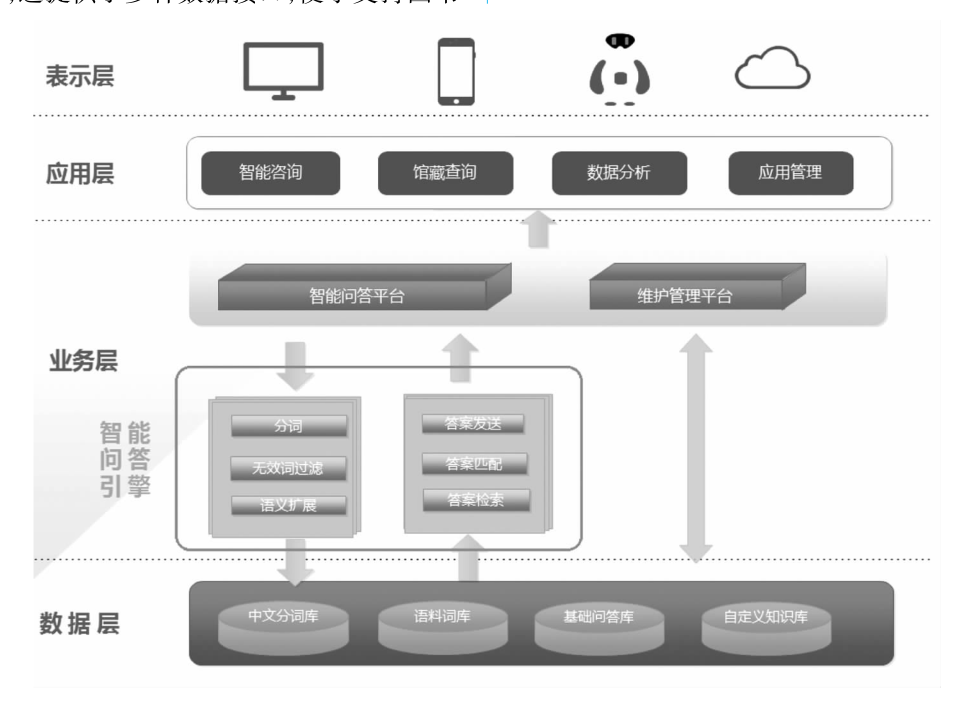


图 2　智能咨询系统整体架构

4.4　基于 Word2vec 的智能问答引擎设计

智能问答引擎的基本技术原理是对语句进行预处理,使用 Word2vec 训练词向量,利用词向量表对用户问题的关键词进行语义扩展,最后采取基于句子共现词的相似度计算实现答案的匹配。

4.4.1　智能问答引擎运行流程

智能问答引擎的运行流程如图 3 所示:当问答引擎接受到用户所提出的问句时,首先依据中文词典和停用词典将用户的问句进行分词、停用词去除等预处理,便得到经过处理的候选词词组。其次,将获取的候

选词词组与训练完成后的词向量库进行比较,取出与特征词组相似度高的若干个词作为语义扩展后的特征词组。最后,使用扩展后的特征词组与已构建的问答知识库中问题进行基于句子共现词相似度的匹配。选择匹配值最高的答案返回给用户,而对于无匹配答案则提供人工服务。

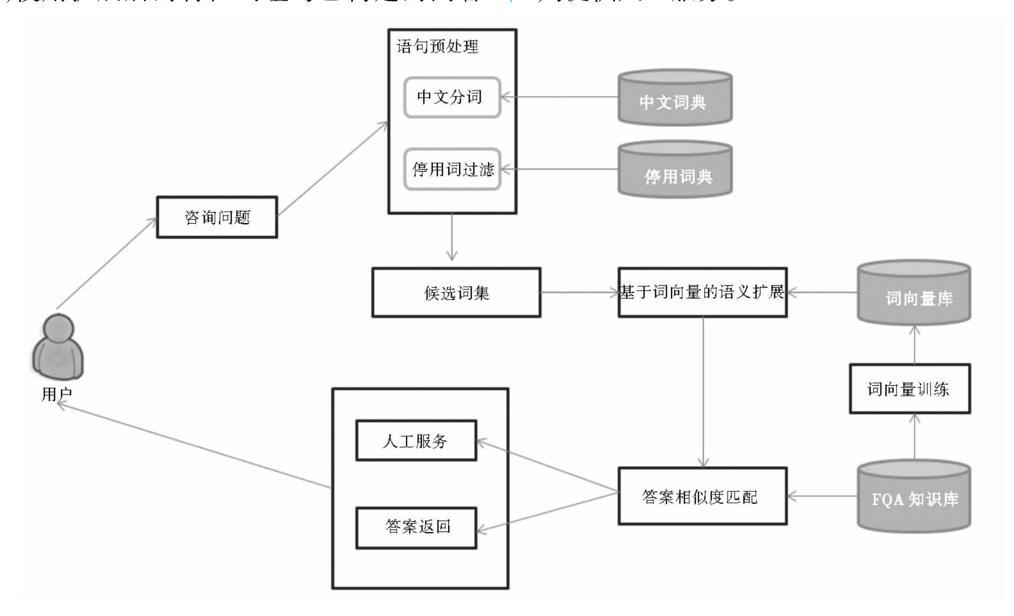


图 3　智能问答引擎运行流程

4.4.2　语句预处理

语句预处理包括对中文进行分词和对停用词进行过滤。对停用词的过滤采用与停用词库进行比对的方式。在这之前,首先需要进行中文分词处理,将用户的问句转为有效的词的表示。本系统应用了基于词典分词的最大匹配分词算法。最大匹配分词的扩展主要有正向最大匹配和逆向最大匹配两种算法[14]。经过比较分析,由于英文单词间是以空格来进行分隔的,所以使用正向最大匹配算法对英文进行分词的效率和词表命中的准确率较高。由于中文词汇结构复杂,使用逆向最大匹配算法会更加准确,因此本系统中选择逆向最大匹配算法(见图 4)。

逆向最大匹配算法在智能咨询引擎中进行分词的过程为:当待分语句为 S1 时,首先设定最大切词词长"MaxLen 值"为 m,m 的设定为分词词典中有效词的最大长度,从右向左取待分语句 S1 中的 m 个字作为候选字串记为"W",查找已有词典对"W"进行匹配。如果匹配成功,则将该字串作为一个词输出到分词结果集"S2"中。如果匹配不成功,则将该字段最右边的一个字去掉,将剩余的字作为新字串重新进行匹配,直到所有词都切分完成,最后输出分词结果集"S2"。

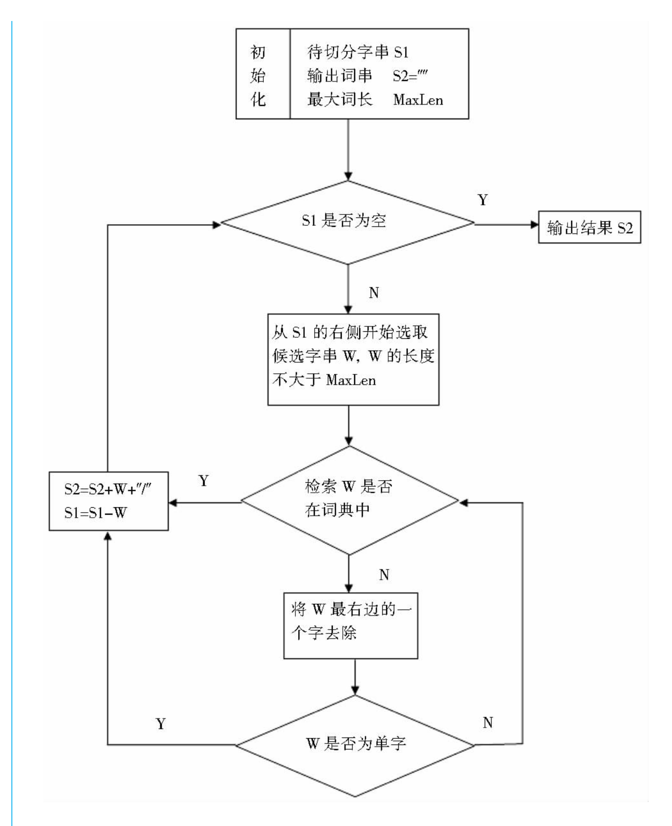


图 4　逆向最大匹配算法

4.4.3　基于 Word2vec 的词向量训练和语义扩展

训练样本库首先要对词向量的训练模型进行选择,Word2vec 主要有两种训练模型即连续词袋模型(CBOW)和跳字模型(Skip-gram)[15]。CBOW 模型是根据周围词预测中心词,再根据中心词的预测结果情况,利用 GradientDesent 方法调整周围词的向量,从而获得整个文本里面所有词的词向量。而 Skip-gram 模

型则是根据中心词来预测周围词,使用周围词的预测情况来调整中心词的词向量,需要文本中所有的字进行处理。从训练模式可以看出 CBOW 的训练效率更高,但在语义分析方面的准确率不如 Skip-gram 模型[16]。以"如何 能 借 图书馆 的 书"这句话的三元词组为例,按照词顺序连续训练的方式只能得到 4 个三元词组 "如何能借"" 能借图书馆""借图书馆的""图书馆的书",这句话本身表达的意思是"如何借书",但是这 4 个三元词组都没有准确的表达出来句子的意思,而使用 Skip-Gram 模型允其跳字,即可以使用不相邻词组成多个三元词组,如表 1 所示:

表 1 Skip-Gram 训练词集示例

Skip-Gram 训练
"如何能借""如何能图书馆""如何能的""如何能书""能借图书馆""能借的""能借书""借图书馆的""借图书馆书""借的书""图书馆的书""如何借图书馆""如何借的""如何借书""能借图书馆""能借的""能借书""图书馆的书"

由表 1 可以看出,当使用 Skip-Gram 模型进行语料训练的时候,能够覆盖到全部的语义组合,实际要表达的意义"如何借书"正在其中,词向量也更加能够反映出真正的文本语义。因此 Skip-gram 模型更加适合于语义分析,所以在本系统中选择 Skip-Gram 模型为词向量的训练模型。Skip-Gram 模型的词向量训练的数学公式可以表示为:

$$\prod_{t=1}^{T}\prod_{-m\leqslant j\leqslant m,j=0}P(\omega^{(t+j)}|\omega^{(t)}) \quad \text{公式(1)}$$

在公式(1)中,T 表示窗口中心词的位置,m 表示的是滑动窗口的大小。在实际的训练中,以"如何 能 借 图书馆 的 书"为例,假设计算到"借"这个词语,当设置 m 值为 2 时,针对"借"这个词,需要分别计算在这个词与相邻的前两个和后两个词的概率,有 P(如何|借)、P(能|借)、P(图书馆|借)、P(的|借)。由此可以看出,如果滑动窗口选择太大或过小都会影响模型训练的效果。经过对样本库的测试后,在实际训练中将滑动窗口值设定为 4。对关键词语义扩展的具体做法是:在训练后的词向量表中查询与关键词余弦值接近的词,将比较阈值设置为 0.8,若大于该值判断为相似词,将此词作为查询词的扩展词。在实际计算中发现高于这个值词的可能很多,考虑到系统计算精度,选取相似度最高的前 3 个词为当前词的扩展词,放入扩展后的词集中,为后续的问答匹配做准备。

4.4.4 基于共现词相似度的答案匹配

智能咨询引擎在答案匹配的方法是:将扩展后的词集与知识库中问题的问句进行基于共现词的相似度计算,选择相似度最高的作为匹配问句,再从 FAQ 数据库中根据问句查询相应答案并返回。基于共现词相似度匹配算法文的基本原理就是比较的两个语句中共现词汇的数量,数量越多则代表这两个语句的相似度也越高。相似度计算公式可以表示为:

$$\text{Similarity}(S_i,S_j)=\frac{|\{w_k|w_k\in S_i\cap w_k\in S_j\}|}{\log(|S_i|)+\log(|S_j|)} \quad \text{公式(2)}$$

公式(2)中,Si、Sj 表示需要比较的两个句子,Wk 表示句子中的词,分子表示同时出现在两个句子中的相同词的个数。分母取对数,是为了抵消长度相差较大句子比较时对计算值的影响。

4.5 维护和管理平台设计

维护和管理平台主要包括问答管理、系统管理、数据统计和知识库管理等 4 个子模块。问答管理流程的设计使用了协同办公的理念,增加了问题分配和问答审核机制。具体的做法是:由值班馆员根据用户提交问题的类型,将问题分配给相关部门的值班人员进行回答,由各部门的负责人员审核后对读者进行回复,同时更新知识库。系统管理模块包括用户管理和登陆日志管理两部分,用户管理中采用分级角色管理:管理员、学科管理员、学科馆员、运营维护员,支持不同级别权限用户的查询、添加、编辑和删除等操作。数据统计模块是对智能咨询系统运行和用户行为的统计和展示,可根据不同维度对数据进行统计分析和可视化展示。知识库管理包括对新知识库整体的导入与导出,对已有知识库的语料的添加、修改和删除操作。为保障数据安全,仅限学科管理员能够对语料进行审核入库和删除等操作,学科馆员只能进行编辑和查看语料等操作。

5 系统实现

根据上述提出的系统设计与算法,本节实现了图书馆智能咨询系统。该系统环境配置如下:

编程语言:前端页面使用 vue + react,核心算法用 C + + 语言。

数据库:MYSQL 5.6。

运行环境:服务器操作系统使用 Windows2008, Web 服务器使用 Tomcat 5.5。

5.1 智能问答引擎的实现

5.1.1 中文分词

通过使用逆向最大匹配算法将引擎获得的问句进行中文分词处理,系统选用 jieba 分词词库和自建的图

张乐. 词向量语义扩展技术在图书馆智能咨询系统的应用与实现[J]. 图书情报工作,2020,64(18):126－136.

书馆 FAQ 分词词库,输出为切分好的词串。具体伪代码如下:

```
Vector WordSegment(String sentence, Dict wordList)
var  maxLen = 7 //最大词组的长度
var  result //输出词串
var  index = 0
while (sentence.length() > 0){
  var word = sentence[index:maxLen]
  while(1){ // 内循环
    if (wordList.find(word)) { // 查词典,看 word 是否在词典中
      result.append(word)
      index = index + word.length() // 更新游标
      break //跳出内循环
    }
    else {                  // word 不在词典中
      if (word.length() = = 1) { //只剩一个字
        result.append(word)
        index = index + word.length()
        break //跳出内循环
      }
      word = word.pop_back(),//去掉最右侧一个字
    }

  sentence = sentence[index:] // 将匹配到的词从 sentence 左侧去
掉
}
return result //返回结果[word1, word2, word3, ...]
```

5.1.2 停用词过滤

对中文进行分词后得到的词串,通过停用词过滤将不需要的词去除,输出过滤后的有效词串。

停用词表使用百度停用词表和自定义词表,输出为切分后的词串 wordList。具体伪代码如下:

```
Vector StopWordsFilter(Vector words, Dict stopWordList)
Var  result
var  index = 0
while (index < words.length()){
  var word = words[index]
  index + = 1
  if (stopWordList.find(word)) { // 在停用词典内
    continue //跳过
  }
  result.append(word) // 未在停用词典内
}
return result //返回结果[word1, word2, word3, ...]
```

5.1.3 基于 Word2vec 的词向量训练

系统采用 python 的 gensim 为训练工具,使用东南大学图书馆常用 FAQ 知识库、图书馆设施和规章知识库为训练样本。关键参数设置为:

-train = trainfile;-output = FAQ.vec; -cbow = 0(训练模型选择:Skip-Gram);-size = 200(向量维度);-window = 4(滑动窗口)。部分词向量训练结果如图 5 所示:

```
图谋 -0.080328904 -0.38751355 0.4338339 0.22465171 -0.7884939 -0.32396036 0.1053
图片 2.1176443 2.3692298 -0.36268267 -0.32863328 -2.9598074 -1.0689933 2.748145
图片社 0.21411207 0.586781 0.46677858 -1.4756463 0.024095818 0.73342806 0.25909
图谱 0.822712 1.3480308 -1.842866 -0.20504187 -0.4646954 -0.41304004 -0.11278636
图示 0.95335937 0.10484978 -1.4966564 0.8941801 -1.0356681 0.37732416 0.778524
图式 0.088620126 0.041640364 0.007172336 0.014298022 -0.10763017 0.04662903 0.1
图书 -1.2392286 3.3644757 2.5200496 -0.15047817 -0.73181665 -2.043708 0.4544578
图书馆 -0.10228585 1.4506063 2.3656974 2.151757 3.0822504 -2.6789072 3.2704575 0
图书室 -0.20917505 0.076000966 0.14207293 -0.1392632 0.04943845 -0.117100164 0.0
图说 -0.024342665 0.022511588 -0.04546367 -0.05530472 0.012335857 0.0079244515
图腾 0.70613045 0.06461371 2.377354 -1.1895387 0.55518717 2.5359783 -0.5834744
图文 1.3509521 1.9034424 1.9373567 -1.7429898 -2.1877098 2.178882 0.6168056 -0.0
图象 1.181109 0.32476628 -0.15417813 -0.08778427 -0.46677792 -0.2043813 0.35229
图像 2.8962748 0.81774235 -1.2575828 -0.6568581 -0.837637 0.0426858 1.2458569 0.
图形 5.9158764 0.84704036 -1.1898754 1.8543832 -4.638064 2.2729003 2.816778 -0.3
图样 1.2262474 0.9116692 -0.19078358 -0.59467113 -0.61900425 0.7874932 -0.69735
图章 1.0161691 1.4710507 0.538855 0.19301178 -0.30157813 0.88103443 -0.29140043
图纸 1.5503052 2.0734816 0.28666842 1.7776623 -0.8467179 -0.15986298 0.10136588
```

图 5 部分词向量训练结果

训练完成后,以“延期”为输入词,通过和训练过的词向量库进行比较计算,得出和“延期”相近的词的集合以及相似度数值。如图 6 所示:



图 6 “延期”相似词计算

5.1.4 基于词向量的语义扩展

通过训练好的词向量库对预处理过的有效词串进行语义扩展,取与输入词相似度最高的前 3 个词为当前词的扩展词,具体伪代码如下:

```
Vector WordFilter(Vector words,Model word2VecModel)
Var  result
Var  index = 0
var  topn = 3       //取 top3 词向量的扩展
while (index < words.length()){
  var word = words[index]
  index + = 1
  w2v = word2VecModel.most_similar(word,topn)

  result.append(word) // 原词
  for (w in w2v) { // 在停用词典内
```

```
    result. append(w) //语义扩展词
  }
}
return result  //返回结果[word1, word2, word3, ...]
```

其中，输入：words 为停用词过滤后的词，word2VecModel 为词向量模型。输出：语义扩展后的词串。

5.1.5 基于共现词的句子相似度匹配

根据相似度算法，进行句子的相似度匹配，计算输出相似度值，具体算法伪代码如下：

```
double SentenceSimilarity(String sen1,String sen2)
counter = 0
for word in sen1:  //共现词计算
    if word in sen2:
        counter + = 1
double similarity = counter/(log(len(sen1)) + log(len(sen2))) //相似度计算
return similarity
```

5.2 用户端和管理端的实现

用户端提供了网页版和微信版服务。智能咨询系统支持两种类型的用户使用，即匿名咨询和认证用户咨询，其中认证用户可以获得人工服务，通过微信端提供了账户绑定服务，只需要一次绑定微信号和用户的学号后即可免登陆，可以方便地留言提问，得到人工回复。微信版的用户界面和人工服务交互界面如图 7 所示：



图 7 用户界面和人工服务交互界面

智能分析和管理平台为图书馆相关人员提供了基于网页端的管理界面，包括问答管理、站点数据分析、知识库管理以及系统管理等功能。整体实现页面如图 8 所示：

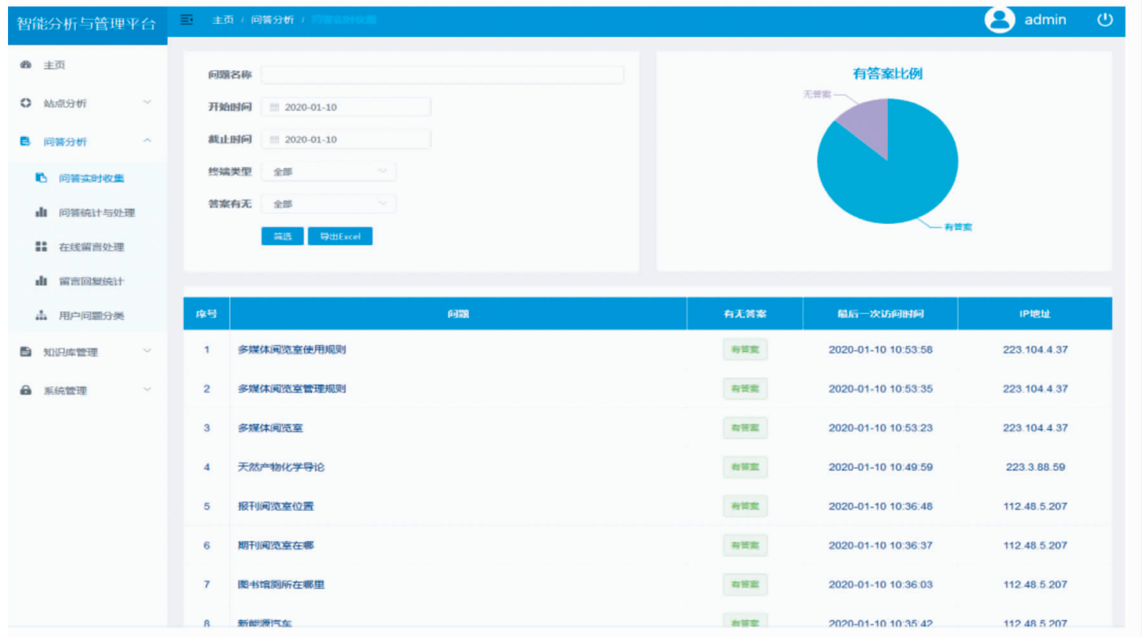


图 8 分析和管理平台界面

由于篇幅有限,以下仅根据4.5节的设计展示问答管理中的人工服务以及入库操作的不同角色馆员协同工作的流程的实现。

当用户提交问题“学校餐厅对外开放吗?”,智能咨询引擎匹配无答案后,学科馆员在后台获取到分配的未回答问题。点击相关问题,系统跳转至“在线留言处理”,对问题进行回答操作。如图9所示:



图9　人工服务问答编辑操作界面

答案推送至用户后,系统将问题、答案及操作馆员的信息推送给学科管理员,并进入待审核入库状态。学科管理员审核确定问答是否达到入库的标准,符合标准的执行入库操作,不符合的则驳回删除,如图10所示:



图10　问答审核界面

学科管理员将问答审核入库后,当智能咨询系统再次收到用户相同的问题时,即可由智能问答引擎系统直接给出答案。如图11所示:

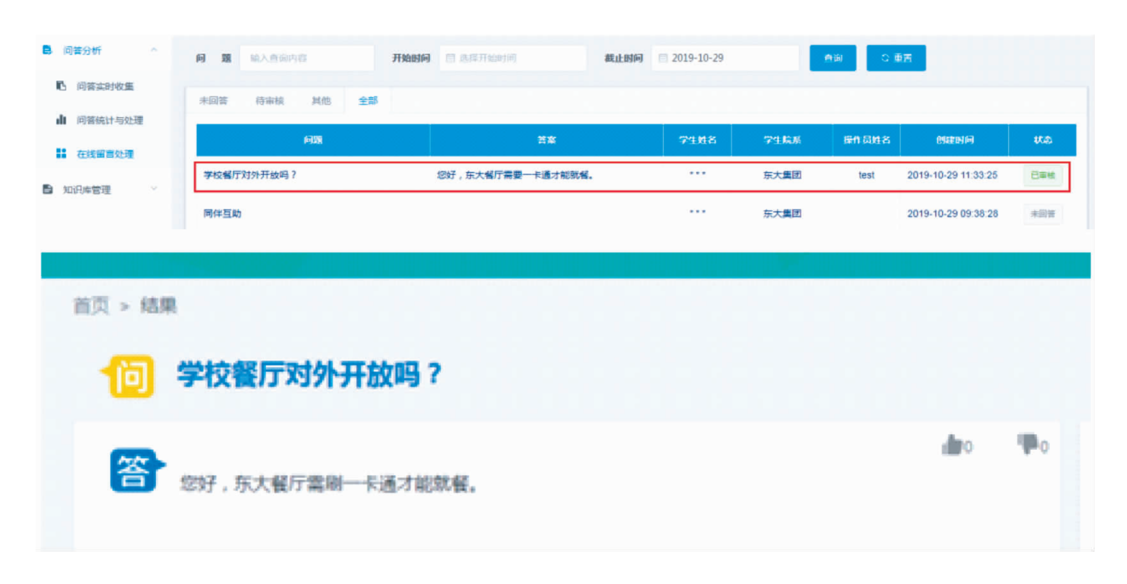


图11　问答入库和自动回答反馈

6 图书馆智能咨询系统的运行情况

东南大学图书馆智能咨询系统自 2019 年 10 月开始上线试用,至 2020 年 1 月累计运行 3 个月,以下通过对系统数据的统计,分析使用情况。

6.1 用户使用情况分析

智能咨询系统累计总访问量为 4 634 人次,查询 4 420个问题。平均日均访问量约为 51 人次/天。设定工作时间为 8:30 – 17:00,其余时间为非工作时间。

由图 12 可见,在非工作时间访问智能咨询系统的用户占总访问人数的 29.3%,用户咨询问题占总咨询数的 31.1%,说明东南大学图书馆的用户在非工作时间对图书馆信息咨询服务的需求也十分明显。

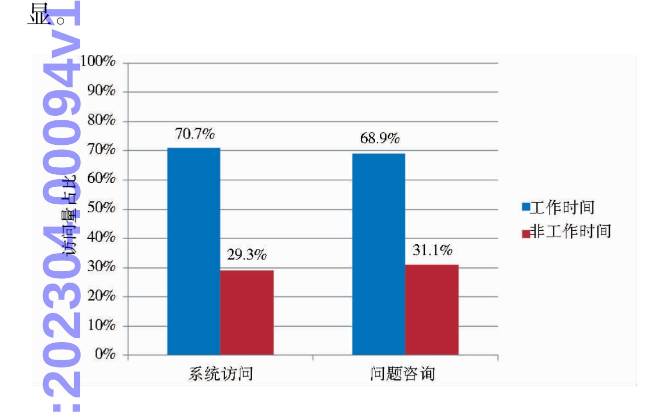


图 12 用户使用咨询系统时间分布

在咨询内容方面,关于馆舍规章等常规问题内容的约占问题的 51%,关于馆藏书目的检索约占 36%。与此同时,使用东南大学图书馆公众号以及线上工具咨询的人工服务的回复数量有较大的下降,尤其是对图书馆常规问题的咨询。如图 13 所示:

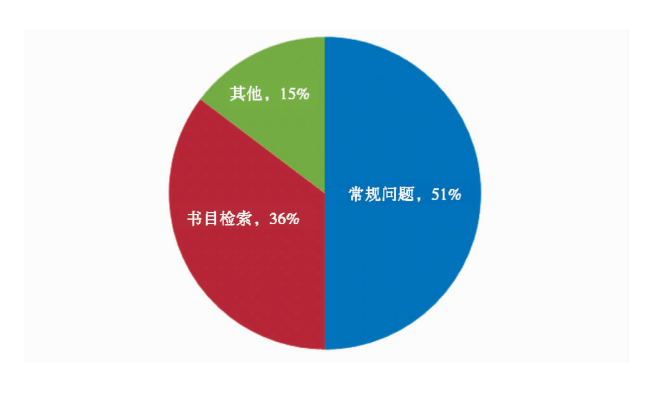


图 13 咨询内容分布

综上所述,图书馆智能咨询系统的使用在延长图书馆信息咨询服务时间和降低馆员劳动量方面的效果较为明显。

6.2 图书馆智能咨询系统运行效果

在早期系统测试阶段,出现有些同义词无匹配的情况,经过测试发现,是由于基于 Skip-gram 模型词向量的训练对语料的数量较为敏感,以及初始的匹配阙值设置过高造成的。通过降低阙值以及语料库的累加后,获得了较为理想的匹配效果,同时应用推送相似问题链接的方法来提升咨询效果和用户友好度。以问题“图书超期归还规则”为例,使用“我的图书逾期”“逾期归还”“超期怎么办”“图书过期”等相似语义的问题进行咨询,均可获得“图书超期归还”的相关图书馆规则回复,同时将相似问题推送给用户。见图 14。

在系统上线运行过程中,共收到 1 927 个信息咨询问题,智能咨询引擎自动回复了 1 436 个问题,约为 74.5%,有 99 个没有回答的问题通过人工服务的方式进行了提交后台处理。相对于图书馆目前使用的基于关键词的自动问答机器人不到 50% 的问题应答率有较大水平的提高。经过对未能自动回复问题的分析,发现主要原因是系统目前只接入图书馆领域的知识库,而用户咨询的问题超出范围所致。随着知识库内容的不断扩充,智能功能问答引擎的应答率应能够有进一步的提升。

6.3 维护与管理平台的使用情况

在系统实际使用中,协同工作的效果较为理想,当值班馆员收到人工服务请求后,通过管理系统快速将相关问题转发给相关负责馆员处理,减少了沟通的时间,在降低了回复延时的同时提高了回复质量。对于知识库的扩展方面,基于角色的分级管理模式以及入库审核机制的设立提高了知识库入库问题的质量和安全性。

7 结语

东南大学图书馆通过对基于 Word2vec 词向量语义扩展技术的研究和使用,很好地解决了自动问答系统在语义扩展方面存在的缺陷,实现了智能化的图书馆信息咨询系统。智能咨询系统在延长咨询服务时间、提高咨询效果、降低咨询馆员工作量和馆员协同工作方面较好地满足了图书馆信息咨询需求,但也有一些不足之处,下一步将从以下 3 方面加强系统建设:①提供更多种的咨询服务终端和功能如接入语音识别功能,开发微信小程序客户端,接入实体机器人系统



图 14　咨询效果测试

等,进一步提升用户使用的体验感。②加强系统知识库建设,可以通过网络下载或者接入的方式为智能问答系统提供除图书馆领域以外知识库的扩展,如聊天库等,但与此同时也要注意知识库内容的审核和管理。③加强对人工智能方面热门技术的研究,如结合使用深度学习和词性标注等技术来主动判断读者意图等,进一步提升图书馆咨询系统的智能化水平。

参考文献:

[1] 郑红京. 近十年中美图书馆参考咨询服务比较研究[J], 高校图书馆工作,2012(6):75－78.

[2] 刘青华,谭红英. 国外参考咨询服务形式浅谈[J],情报科学,2002(6):590－594.

[3] 郭亚军,孟嘉,胡雅悦. 中美一流大学图书馆移动服务比较研究[J]. 图书情报工作,2019,63(11):43－51.

[4] 姚飞, 纪磊, 张成昱, 等. 实时虚拟参考咨询服务新尝试－清华大学图书馆智能聊天机器人[J]. 数据分析与知识发现,2011(4):77－81.

[5] 孙翌,李鲍,曲建峰. 图书馆智能化 IM 咨询机器人的设计与实现[J]. 现代图书情报技术,2011(5):88－92.

[6] 沈奎林, 邵波, 赵华. 利用微信构建图书馆智能问答系统[J]. 图书馆学研究,2015(8):75－80.

[7] 张长恒,黄芳. 利用微信公众平台构建高校图书馆 APP 的技术实现[J]. 图书情报工作,2015,59(4):37－43.

[8] 李丹. 图书馆微信平台建设实践与思考[J]. 现代图书情报技术,2016(4):104－110.

[9] 韩如冰, 叶得学. 问答系统的汉语分词算法研究[J]. 数字技术与应用,2012(5): 114－115.

[10] 欧石燕,胡珊,张帅. 本体与关联数据驱动的图书馆信息资源语义整合方法及其测评[J]. 图书情报工作,2014,58(2):5－13.

[11] 杜文华. 本体的构建及其在数字图书馆中的应用研究[D]. 武汉:武汉大学,2005.

[12] MCCORMICK C. Applying word2vec to recommenders and advertising[EB/OL]. [2020－01－18]. http://mccormickml.com/2018/06/15/applying-word2vec-to-recommenders-and-advertising.

[13] 刘宸毛,琦李彦,朱晓芒,等. 智能应答系统在高校信息化服务中的应用研究[J]. 中国教育信息化,2019(3):43－45.

[14] 周程远,朱敏,杨云. 基于词典的中文分词算法研究[J]. 计算机与数字工程, 2009(3):68－71,87.

[15] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. [2020－01－18]. https://arxiv.org/abs/1301.3781.

[16] 徐成章. 基于 Word2vec 的中文 Web 智能问答系统的研究与设计[D]. 成都:电子科技大学, 2018.

Application and Implementation of Word Vector Semantic Extension Technology in Library Intelligent Consulting System

Zhang Le
Southeast University Library, Nanjing 211189

Abstract: [Purpose/significance] Aiming at the problem of semantic extension in automatic question answering system, this paper proposes a semantic extension technology based on word vector, and designs and implements a library intelligent consulting system. [Method/process] Using word2vec word vector semantic extension technology, Chinese word segmentation and co-occurrence matching technology, an intelligent Q & A engine was designed. Combined with the concept of collaborative office, the library intelligent consulting system was realized, and the operation data of the system was statistically analyzed. [Result/conclusion] The system meets the design requirements well in terms of working hours, consultation effect and back-stage management, and provides reference for the construction of library intelligent consultation system.

Keywords: word vector semantic extension Word2vec intelligent consultation library